DANIEL J. MULLER, SBN 193396
dmuller@venturahersey.com
VENTURA HERSEY & MULLER, LLP
1506 Hamilton Avenue
San Jose, California 95125
Telephone: (408) 512-3022
Facsimile: (408) 512-3023

Attorneys for Plaintiff and the Class

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA – SAN FRANCISCO DIVISION

MICHAEL CHABON, DAVID HENRY HWANG, MATTHEW KLAM, RACHEL LOUISE SNYDER, AND AYELET WALDMAN,

individually and on behalf of all others similarly situated,

Plaintiffs,

v.

OPENAI, INC., OPENAI, L.P., OPENAI OPCO, LLC, OPENAI GP LLC, OPENAI STARTUP FUND GP I, LLC, OPENAI STARTUP FUND I, LP, and OPENAI STARTUP FUND MANAGEMENT, LLC,

Defendants.

Case No.

CLASS ACTION COMPLAINT

CLASS ACTION

JURY TRIAL DEMANDED

Plaintiffs Michael Chabon, David Henry Hwang, Matthew Klam, Rachel Louise Snyder, and Ayelet Waldman ("Plaintiffs"), individually and on behalf of all others similarly situated, bring this action against Defendants OpenAI, Inc., OpenAI, LP, OpenAI OpCo, LLC, OpenAI GP LLC, OpenAI Startup Fund I, LP, OpenAI Startup Fund GP I, LLC, and OpenAI Startup Fund Management, LLC (collectively, "Defendants" or "OpenAI"). Plaintiffs allege as follows based upon personal knowledge as to themselves and their own acts, and upon information and belief as to all other matters:

NATURE OF ACTION

1. This is a class action lawsuit brought by Plaintiffs on behalf of themselves and a Class of authors holding copyrights in their published works arising from OpenAI's clear infringement of their intellectual property.

2. OpenAI is a research company specializing in the development of artificial intelligence ("AI") products, such as ChatGPT.

3. ChatGPT is an AI chatbot, which produces responses to users' text queries or prompts in a way that mimics human conversation.

4. ChatGPT relies on other OpenAI products to function, namely Generative Pre-trained Transformer ("GPT") models. "Generative," in GPT, represents the model's ability to respond to text inquiries, while "Pre-trained" refers to the model's use of training datasets to program its responses, and "Transformer" concerns the model's underlying algorithm allowing it to function.

5. OpenAI has released five versions of GPT models, and the current version of ChatGPT runs on GPT-3.5 and GPT-4, depending on whether the user has subscribed to the premium version of ChatGPT. Only the version of ChatGPT that runs on GPT-3.5 is available at no cost to the public.

6. OpenAI's GPT models are types of "large language model," which is a form of deep-learning algorithm programmed through "training datasets," consisting of massive amounts of text data copied from the internet by OpenAI. The GPT models extract information from their training datasets in order to learn the statistical relationships between words, phrases,

and sentences, which allow them to generate coherent and contextually relevant responses to user prompts or queries.

7. A large language model's responses to user prompts or queries are entirely and uniquely dependent on the text contained in its training dataset, necessarily processing and analyzing the information contained in its training dataset to generate responses.

8. OpenAI incorporated Plaintiffs' and Class members' copyrighted works in datasets used to train its GPT models powering its ChatGPT product. Indeed, when ChatGPT is prompted, it generates not only summaries, but in-depth analyses of the themes present in Plaintiffs' copyrighted works, which is only possible if the underlying GPT model was trained using Plaintiffs' works.

9. Plaintiffs and Class members did not consent to the use of their copyrighted works as training material for GPT models or for use with ChatGPT.

10. Defendants, by and through their operation of ChatGPT, benefit commercially and profit handsomely from their unauthorized and illegal use of Plaintiffs' and Class members' copyrighted works.

JURISDICTION AND VENUE

11. This Court has subject matter jurisdiction of this action pursuant to 28 U.S.C. § 1331 because this case arises under the Copyright Act (17 U.S.C. § 501) and the Digital Millennium Copyright Act (17 U.S.C. § 1202).

12. This Court has personal jurisdiction over Defendants pursuant to 18 U.S.C. §§ 1965(b) & (d), because they maintain their principal places of business in, and are thus residents of, this judicial district, maintain minimum contacts with the United States, this judicial district, and this State, and they intentionally avail themselves of the laws of the United States and this state by conducting a substantial amount of business in California. For these same reasons, venue properly lies in this District pursuant to 28 U.S.C. §§ 1391(a), (b) and (c).

PARTIES

A. Plaintiffs

13. Plaintiff Michael Chabon ("Plaintiff Chabon") is a resident of California. Plaintiff Chabon is an author who owns registered copyrights in many works, including but not limited to, *The Mysteries of Pittsburgh, Wonder Boys, The Amazing Adventures of Kavalier & Clay, the Yiddish Policemen's Union, Gentlemen of the Road, Telegraph Avenue, Fight of the Century, Kingdom of Olive and Ash,* and *Moonglow.* Plaintiff Chabon is the recipient of the Pulitzer Prize for Fiction, Hugo, Nebula, Los Angeles Times Book Prize, and the National Jewish Book Award, among many others achieved over the span of a writing career spanning more than 30 years. Plaintiff Chabon's works include copyright-management information that provides information about the copyrighted work, including the title of the work, its ISBN or copyright registration number, the name of the author, and the year of publication.

14. Plaintiff David Henry Hwang ("Plaintiff Hwang") is a resident of New York. Plaintiff Hwang is a playwright and screenwriter who owns registered copyrights in many works, including but not limited to, *M. Butterfly, Chinglish, Yellow Face, the Dance and the Railroad*, and *FOB*, as well as the Broadway musical*, Flower Drum Song* (2002 revival). Plaintiff Hwang is a Tony Award winner and three-time nominee, a Grammy Award winner who has been twice nominated, a three-time OBIE Award winner, and a three-time finalist for the Pulitzer Prize in Drama. Plaintiff Hwang's works include copyright-management information that provides information about the copyrighted work, including the title of the work, its ISBN or copyright registration number, the name of the author, and the year of publication.

15. Plaintiff Matthew Klam ("Plaintiff Klam") is a resident of Washington D.C. Plaintiff Klam is an author who owns registered copyrights in several works, including but not limited to, *Who is Rich?,* and *Sam the Cat and Other Stories*. Plaintiff Klam is a recipient of a Guggenheim Fellowship, a Robert Bingham/PEN Award, a Whiting Writer's Award, and a National Endowment of the Arts. Plaintiff Klam's works have been selected as Notable Books of the year by *The New York Times, The Los Angeles Times, the Kansas City Star, and the*

Washington Post. Plaintiff Klam's works include copyright-management information that provides information about the copyrighted work, including the title of the work, its ISBN or copyright registration number, the name of the author, and the year of publication.

16. Plaintiff Rachel Louise Snyder ("Plaintiff Snyder") is a resident of Washington, D.C. Plaintiff Snyder is an author who owns registered copyrights in many works, including but not limited to, *Women We Buried, Women We Burned*, *No Visible Bruises – What We Don't Know About Domestic Violence Can Kill Us*, *What We've Lost is Nothing*, and *Fugitive Denim: A Moving Story of People and Pants in the Borderless World of Global Trade*. Plaintiff Snyder is a Guggenheim fellow and the recipient of the J. Anthony Lukas Work-in-Progress Award, the Hillman Prize, and the Helen Bernstein Book Award, and was a finalist for the National Book Critics Circle Award, *Los Angeles Times* Book Prize, and Kirkus Award. Her work has appeared in *The New Yorker*, *The New York Times*, *Slate*, and in many other publications. Plaintiff Snyder's works include copyright-management information that provides information about the copyrighted work, including the title of the work, its ISBN or copyright registration number, the name of the author, and the year of publication.

17. Plaintiff Ayelet Waldman ("Plaintiff Waldman") is a resident of California. Plaintiff Waldman is an author and screen and television writer who owns registered copyrights in several works, including but not limited to, *Love and other Impossible Pursuits*, *Red Hook Road, Love and Treasure*, *Bad Mother, Daughter's Keeper, A Really Good Day, Fight of the Century, and Kingdom of Olives and Ash*. Plaintiff Waldman has been nominated for an Emmy and a Golden Globe and is the recipient of numerous awards including a Peabody, AFI award, and a Pen Award, among others. Plaintiff Waldman's works include copyright-management information that provides information about the copyrighted work, including the title of the work, its ISBN or copyright registration number, the name of the author, and the year of publication.

18. At all times relevant hereto, Plaintiffs have been and remain the holders of the exclusive rights under the Copyright Act of 1976 (17 U.S.C. §§ 101, *et seq.* and all amendments

thereto) to reproduce, distribute, display, or license the reproduction, distribution, and/or display the works identified in paragraphs 13-17, *supra*.

B. Defendants

19. Defendant OpenAI, Inc. is a Delaware nonprofit corporation with its principal place of business located at 3180 18th St., San Francisco, CA 94110.

20. Defendant OpenAI, LP is a Delaware limited partnership with its principal place of business located at 3180 18th St., San Francisco, CA 94110. OpenAI, LP is a wholly owned subsidiary of OpenAI, Inc. that is operated for profit. OpenAI, Inc. controls OpenAI, LP directly and through the other OpenAI entities.

21. Defendant OpenAI OpCo, LLC is a Delaware limited liability company with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI OpCo, LLC is a wholly owned subsidiary of OpenAI, Inc. that is operated for profit. OpenAI, Inc. controls OpenAI OpCo, LLC directly and through the other OpenAI entities.

22. Defendant OpenAI GP, LLC is a Delaware limited liability company with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI GP, LLC is a general partner of OpenAI, LP. OpenAI GP manages and operates the day-to-day business and affairs of OpenAI, LP. OpenAI GP was aware of the unlawful conduct alleged herein and exercised control over OpenAI, LP throughout the Class Period. OpenAI, Inc. directly controls OpenAI GP.

23. Defendant OpenAI Startup Fund I, LP is a Delaware limited partnership with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI Startup Fund I, LP was instrumental in the foundation of OpenAI, LP, including the creation of its business strategy and providing initial funding. OpenAI Startup Fund I was aware of the unlawful conduct alleged herein and exercised control over OpenAI, LP throughout the Class Period.

24. Defendant OpenAI Startup Fund GP I, LLC is a Delaware limited liability company with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI Startup Fund GP I, LLC is the general partner of OpenAI Startup Fund I.

OpenAI Startup Fund GP I is a party to the unlawful conduct alleged herein. OpenAI Startup Fund GP I manages and operates the day-to-day business and affairs of OpenAI Startup Fund I.

25. Defendant OpenAI Startup Fund Management, LLC is a Delaware limited liability company with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI Startup Fund Management, LLC is a party to the unlawful conduct herein. OpenAI Startup Fund Management was aware of the unlawful conduct alleged herein and exercised control over OpenAI, LP throughout the Class Period.

FACTUAL ALLEGATIONS

A. OpenAI's Artificial Intelligence Products

26. OpenAI researches, develops, releases, and maintains AI products with the intention that its products "benefit all of humanity."[1]

27. ChatGPT is among the products OpenAI has developed, engineered, released, and maintained, which utilizes another OpenAI product, GPT models, to respond to text prompts and queries in a natural, coherent, and fluent way through a web interface.

28. OpenAI has released a series of upgrades to its GPT model, including GPT-1 (released June 2018), GPT-2 (February 2019), GPT-3 (May 2020), GPT-3.5 (March 2022), and most recently, GPT-4 (March 2023)[2].

29. The current version of ChatGPT utilizes both GPT-3.5 and GPT-4; however, the version of ChatGPT that allows users to choose between using GPT-3.5 and GPT-4 is only available to subscribers at a cost of $20 per month. Otherwise, users are only able to access the version of ChatGPT that relies on the GPT-3.5 model.[3]

30. OpenAI makes ChatGPT available to software developers through an application-programming interface ("API"), which allows developers to write software

[1] *About*, OpenAI, https://openai.com/about
[2] Fawad Ali, *GPT-1 to GPT-4: Each of OpenAI's GPT Models Explained and Compared*, Make Use Of (Apr. 11, 2023) https://www.makeuseof.com/gpt-models-explained-and-compared/
[3] *Introducing ChatGPT Plus*, OpenAI (Feb. 1, 2023) https://openai.com/blog/chatgpt-plus

programs that exchange data with ChatGPT.[4] OpenAI charges developers for access to ChatGPT by the API on the basis of usage.

B. OpenAI Uses Copyrighted Works in its Training Datasets

31. As mentioned in paragraph 6, *supra*, OpenAI pre-trains its GPT models using a dataset consisting of various sources and content types, including books, plays, articles, and webpage and other written works, to respond accurately to users' prompts and queries.

32. OpenAI has admitted that, of all sources and content types that can be used to train the GPT models, written works, plays and articles are valuable training material because they offer the best examples of high-quality, long form writing and "contain[] long stretches of contiguous text, which allows the generative model to learn to condition on long-range information."[5]

33. Upon information and belief, OpenAI builds the dataset it uses to train its GPT models by scraping the internet for text data.

34. While casting a wide net across the internet to capture the most comprehensive set of content available allows OpenAI to better train its GPT models, this practice necessarily leads OpenAI to capture, download, and copy copyrighted written works, plays and articles.

35. Among the content OpenAI has scraped from the internet to construct its training datasets are Plaintiffs' copyrighted works.

36. In its June 2018 paper introducing the GPT-1 model, *Improving Language Understanding by Generative Pre-Training*, OpenAI revealed that it trained the GPT-1 model using two datasets: "Common Crawl," which is a massive dataset of web pages containing billions of words, and "BookCorpus," which is a collection of "over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance."[6]

[4] *OpenAI API*, OpenAI (June 11, 2020) https://openai.com/blog/openai-api
[5] Alec Radford, *Improving Language Understanding by Generative-Pre-Training*, OpenAI (June 11, 2018).
[6] *Id.*; *see also* Fawad Ali, *GPT-1 to GPT-4: Each of OpenAI's GPT Models Explained and Compared*, Make Use Of (Apr. 11, 2023) https://www.makeuseof.com/gpt-models-explained-and-compared/

37. BookCorpus is a controversial dataset, assembled in 2015 by a team of AI researchers funded by Google and Samsung for the sole purpose of training language models like GPT by copying written works from a website called Smashwords, which hosts self-published novels, making them available to readers at no cost.[7] Despite those novels being largely under copyright, they were copied into the BookCorpus dataset without consent, credit, or compensation to the authors.[8]

38. OpenAI also copied many books while training GPT-3. In the July 2020 paper introducing GPT-3, *Language Models are Few-Shot Learners*, OpenAI disclosed, in addition to using the "Common Crawl" and "WebText" datasets that capture web pages, 16% of the GPT-3 training dataset came from "two internet-based book corpora," which OpenAI simply refers to as "Books1" and "Books2."[9]

39. OpenAI has never revealed what books are part of the Books1 and Books2 datasets or how they were obtained. OpenAI has offered a few clues, admitting that these are internet-based datasets that are much larger than BookCorpus.[10] Based on the figures provided in its GPT-3 introductory paper, Books1 is nine times larger than BookCorpus, meaning it contains roughly 63,000 titles, and Books2 is 42 times larger, meaning it contains about 294,000 titles.[11]

40. A limited number of internet-based book corpora exist that contain this much material, meaning there are only a handful of possible sources OpenAI could have used to train the GPT-3 model.

41. Project Gutenberg is an online archive of e-books whose copyrights have expired. Project Gutenberg has long been popular for training AI systems due to the lack of copyright. In 2018, a team of AI researchers created the "Standardized Project Gutenberg Corpus," which

---

[7] Jack Bandy, *Dirty Secrets of BookCorpus, a Key Dataset in Machine Learning*, Medium (May 12, 2021) https://towardsdatascience.com/dirty-secrets-of-bookcorpus-a-key-dataset-in-machine-learning-6ee2927e8650

[8] *Id.*

[9] Tom B. Brown, *Language Models are Few-Shot Learners*, OpenAI (July 22, 2020).

[10] *Id.* at 9.

[11] *Id.*

contained "more than 50,000 books."[12] On that information and belief, the OpenAI Books1 dataset is based on either the Standardized Project Gutenberg Corpus or Project Gutenberg itself, because of the roughly similar sizes of the two datasets.

42. As for the Books2 dataset, the only "internet-based books corpora" that have ever offered that much material are infamous "shadow library" websites, like Library Genesis ("LibGen"), Z-Library, Sci-Hub, and Bibliotik, which host massive collections of pirated books, research papers, and other text-based materials.[13] The materials aggregated by these websites have also been available in bulk through torrent systems.[14]

43. These illegal shadow libraries have long been of interest to the AI-training community. For instance, an AI training dataset published in December 2020 by EleutherAI called "Books3" includes a recreation of the Bibliotik collection and contains nearly 200,000 books.[15] On information and belief, the OpenAI Books2 dataset includes books copied from these "shadow libraries," because those are the sources of trainable books most similar in nature and size to OpenAI's description of Books2.

44. When OpenAI introduced GPT-4 in March 2023, the introductory paper contained no information about the dataset used to train it.[16] Instead, OpenAI claims that, "[g]iven both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about . . . dataset construction."[17]

45. Regarding GPT-4, OpenAI has conceded that it did filter its dataset "to specifically reduce the quantity of inappropriate erotic text content," implying that it again used a large dataset containing text works.[18]

C. OpenAI Unlawfully Infringed Plaintiffs' Copyrights

---

[12] Martin Gerlach, et al., *A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics*, Cornell University (Dec. 19, 2018), https://arxiv.org/pdf/1812.08092.pdf

[13] *See* Claire Woodcock, *'Shadow Libraries' Are Moving Their Pirated Books to The Dark Web After Fed Crackdowns*, Vice (Nov. 30, 2022).

[14] *Id.*

[15] *See* Alex Perry, *A giant online book collection Meta used to train its AI is gone over copyright issues*, Mashable (Aug. 18, 2023).

[16] *GPT-4 Technical Report*, OpenAI (Mar. 27, 2023).

[17] *Id.* at 2.

[18] *Id.* at 61.

46. As explained, ChatGPT's responses to user queries or prompts, like other large language models, rely on the data upon which it is trained to generate responsive content. For example, if ChatGPT is prompted to generate a writing in the style of a certain author, GPT would generate content based on patterns and connections it learned from analysis of that author's work within its training dataset.

47. On information and belief, the reason ChatGPT can generate a writing in the style of a certain author or accurately summarize a certain copyrighted book and provide in-depth analysis of that book is because it was copied by OpenAI and copied and analyzed by the underlying GPT model as part of its training data.

48. When ChatGPT is prompted to summarize copyrighted written works authored by Plaintiffs, it generates accurate, in-depth summaries and analyses of their works.

49. For example, when prompted, ChatGPT accurately summarized Plaintiff Chabon's novel *The Amazing Adventures of Kavalier & Clay.* When prompted to identify examples of trauma in the *Amazing Adventures of Kavalier & Clay*, ChatGPT identified six specific examples, including how the main character's "experiences in Europe, including witnessing the persecution of Jews and the loss of his family, haunt him throughout the story." When asked to write a paragraph in the style of *The Amazing Adventures of Kavalier & Clay,* ChatGPT generated a passage imitating Plaintiff Chabon's writing style including references to the characters dealing with "the weight of the world at war." *Exhibit A.*

50. ChatGPT similarly provided in depth summaries and analyses of Plaintiff Hwang's play, *The Dance and the Railroad*. For example, when prompted, ChatGPT identified five key themes from *The Dance and the Railroad*, including "art and creativity as a form of resistance" and "using art as a form of escape from the harsh realities and dehumanization of labor." Additionally, when prompted to produce a screenplay in the style of *The Dance and the Railroad*, ChatGPT produced a script written in Plaintiff Hwang's style, which generated a screenplay involving a Chinese laborer toiling on the Central Pacific Railroad that "believe[s] in the power of art to keep [their] spirits alive." *Exhibit B.*

51. Likewise, ChatGPT provided in depth summaries and analyses of Plaintiff Klam's works. For example, when prompted, Chat GPT accurately summarized Plaintiff Klam's novel *Who is Rich?* and correctly analyzed the key relationships between the novel's central character and the other characters in the novel. When asked to identify the main themes in *Who is Rich?* Chat GPT accurately identified seven main themes of the novel including "mid-life crisis and identify." Further, when prompted to write a paragraph in the style of *Who is Rich?*, ChatGPT generated random passages authentically written in Plaintiff Klam's writing style, including a reference to navigating the "treacherous waters of midlife." *Exhibit C.*

52. In the same vein, after being prompted to summarize Plaintiff Snyder's book, *What We've Lost is Nothing*, ChatGPT accurately identified themes included within the novel, such as "safety, perception, and the fragility of human relationships." Similarly, once prompted, ChatGPT accurately analyzed the theme of safety using a specific example from the text of Plaintiff Snyder's copyrighted work, explaining that "the theme of safety is examined through the lens of a series of burglaries that occur in a suburban neighborhood . . . and how these incidents affect the characters and their perceptions of the world around them." ChatGPT was also able to generate random passages authentically written in Plaintiff Snyder's writing style when prompted. *Exhibit D.*

53. Additionally, ChatGPT provided in depth summaries and analyses of Plaintiff Waldman's works. For instance, when prompted to summarize Plaintiff Waldman's novel *Love and Other Impossible Pursuits*, Chat GPT accurately provided a summary and analysis of the novel. When prompted to identify specific instances of grief in *Love and other Impossible Pursuits,* ChatGPT identified five specific instances of grief, including the protagonist Emelia's loss of her infant daughter, a "loss that occurred before the events of the novel and [that] continue to haunt Emelia, affecting her emotional state and relationships." When prompted to write a paragraph in the style of *Love and Other Impossible Pursuits,* ChatGPT generated a paragraph imitating Plaintiff Waldman's writing style, including references to the "weight of her daughter's absence." *Exhibit E.*

54. At no point did ChatGPT reproduce any of the copyright management information Plaintiffs included with their published works.

55. Furthermore, at no point did Plaintiffs authorize OpenAI to download and copy their protected works, as described above.

CLASS ALLEGATIONS

56. Plaintiffs bring this action pursuant to the provisions of Rules 23(a), 23(b)(2), and 23(b)(3) of the Federal Rules of Civil Procedure, on behalf of themselves and the following proposed Class:

> All persons or entities in the United States that own a United States copyright in any written work that OpenAI used to train any GPT model during the Class Period.

57. Excluded from the Class are Defendant, its employees, officers, directors, legal representatives, heirs, successors, wholly- or partly-owned, and its subsidiaries and affiliates; proposed Class counsel and their employees; the judicial officers and associated court staff assigned to this case and their immediate family members; all persons who make a timely election to be excluded from the Class; governmental entities; and the judge to whom this case is assigned and his/her immediate family.

58. This action has been brought and may be properly maintained on behalf of the Class proposed herein under Federal Rule of Civil Procedure 23.

59. Numerosity. Federal Rule of Civil Procedure 23(a)(1): The members of the Class are so numerous and geographically dispersed that individual joinder of all Class members is impracticable. On information and belief, there are at least tens of thousands of members in the Class. The Class members may be easily derived from Defendants' records.

60. Commonality and Predominance. Federal Rule of Civil Procedure 23(a)(2) and 23(b)(3): This action involves common questions of law and fact, which predominate over any questions affecting individual Class members, including, without limitation:

a. Whether Defendants engaged in the conduct alleged herein;

b. Whether Defendants violated the copyrights of Plaintiffs and the Class when they

downloaded and copied Plaintiffs' and the Class's copyrighted books;

c. Whether ChatGPT itself is an infringing derivative work based on Plaintiffs' and the Class's copyrighted books;

d. Whether the text responses of ChatGPT are infringing derivative works based on Plaintiffs' and the Class's copyrighted books;

e. Whether Defendants violated the DMCA by removing copyright-management information from Plaintiffs' and the Class's copyrighted books;

f. Whether Defendants were unjustly enriched by the unlawful conduct alleged herein;

g. Whether Defendants' conduct violates the California Unfair Competition Law;

h. Whether Plaintiffs and the other Class members are entitled to equitable relief, including, but not limited to, restitution or injunctive relief; and

i. Whether Plaintiffs and the other Class members are entitled to damages and other monetary relief and, if so, in what amount.

61. Typicality. Federal Rule of Civil Procedure 23(a)(3): Plaintiffs' claims are typical of the other Class members' claims because, among other things, all Class members were comparably injured through Defendants' wrongful conduct as described above.

62. Adequacy. Federal Rule of Civil Procedure 23(a)(4): Plaintiffs are adequate Class representative because their interests do not conflict with the interests of the other members of the Class they seeks to represent; Plaintiff have retained counsel competent and experienced in complex class action litigation; and Plaintiffs intend to prosecute this action vigorously. The interests of the Class will be fairly and adequately protected by Plaintiffs and their counsel.

63. Declaratory and Injunctive Relief. Federal Rule of Civil Procedure 23(b)(2): Defendants have acted or refused to act on grounds generally applicable to Plaintiffs and the other members of the Class, thereby making appropriate final injunctive relief and declaratory relief with respect to the Class as a whole.

64. Superiority. Federal Rule of Civil Procedure 23(b)(3): A class action is superior

to any other available means for the fair and efficient adjudication of this controversy, and no unusual difficulties are likely to be encountered in the management of this class action. The damages or other financial detriment suffered by Plaintiffs and the other Class members are relatively small compared to the burden and expense that would be required to individually litigate their claims against Defendants, so it would be impracticable for the members of the Class to individually seek redress for Defendants' wrongful conduct. Even if Class members could afford individual litigation, the court system could not. Individualized litigation creates a potential for inconsistent or contradictory judgments, and increases the delay and expense to all parties and the court system. By contrast, the class action device presents far fewer management difficulties, and provides the benefits of single adjudication, economy of scale, and comprehensive supervision by a single court.

CAUSES OF ACTION

FIRST CAUSE OF ACTION

DIRECT COPYRIGHT INFRINGEMENT, 17 U.S.C. § 106, *et seq*.

65. Plaintiffs hereby incorporate by reference the allegations contained in the preceding paragraphs of this Complaint.

66. Plaintiffs bring this claim on behalf of themselvess and on behalf of the Class against Defendants.

67. As the owners of the registered copyrights in books used to train OpenAI's GPT models, Plaintiffs and the Class hold the exclusive rights to those works under 17 U.S.C. § 106.

68. Plaintiffs have obtained copyright registrations for each of the works identified in Exhibit B.

69. On information and belief, to train OpenAI's GPT models, OpenAI relied on harvesting mass quantities of content from the public internet, including Plaintiffs' and the Class's books, which are available in digital formats.

70. Because OpenAI's GPT models cannot function without the expressive information extracted from Plaintiffs' and Class members' works and retained by the GPT

models, GPT and ChatGPT are themselves infringing derivative works without Plaintiffs' and Class members' permission and in violation of their exclusive rights under the Copyright Act.

71. Plaintiffs and the Class never authorized OpenAI to make copies of their written works, make derivative works, publicly display copies (or derivative works), or distribute copies (or derivative works). Each of those rights belong exclusively to Plaintiffs and Class members under copyright law.

72. By and through the actions alleged above, OpenAI has infringed and will continue to infringe Plaintiffs' and the Class's copyrights.

73. OpenAI's acts of copyright infringement have been intentional, willful, and in callous disregard of Plaintiffs' and Class members' rights. OpenAI knew at all relevant times that the datasets it used to train its GPT models contained copyrighted materials, and that its acts were in violation of the terms of use of the materials.

74. OpenAI engaged in the infringing acts described herein for its own commercial benefit.

75. As a direct and proximate result of OpenAI's wrongful conduct, Plaintiffs have been substantially and irreparably injured by OpenAI's acts of direct copyright infringement in an amount not readily capable of determination and, unless permanently enjoined from further acts of infringement and continuing to use and distribute GPT models trained using Plaintiffs' and Class members' copyrighted materials without permission, OpenAI will cause additional irreparable harm for which there is no adequate remedy at law. Plaintiff and the Class are thus entitled to permanent injunctive relief preventing OpenAI from engaging in any further infringement of Plaintiffs' and the Class's copyrighted works.

76. Plaintiffs are further entitled to recover statutory damages, actual damages, restitution of profits, and other remedies provided by law.

SECOND CAUSE OF ACTION

VICARIOUS COPYRIGHT INFRINGEMENT
17 U.S.C. § 106

77. Plaintiffs incorporate by reference all allegations of the preceding paragraphs as though fully set forth herein.

78. Plaintiffs bring this claim on behalf of herself and on behalf of the Class against Defendants.

79. Defendant OpenAI, LP is the for-profit subsidiary of Defendant OpenAI, Inc. and is principally responsible for and dedicated to the development of the GPT models and ChatGPT products at issue in this action. Defendant OpenAI Startup Fund Management, LLC exercised control over Defendant OpenAI, LP, along with Defendant OpenAI GP, LLC, which is the general partner of Defendant OpenAI, LP, responsible for managing and operating the day-to-day business affairs of Defendant OpenAI, LP, and is wholly owned and controlled by Defendant OpenAI, Inc., along with Defendant OpenAI OpCo, LLC. Upon information and belief, Defendant OpenAI Startup Fund I, LP played a vital role in the foundation of Defendant OpenAI, LP, including providing initial funding and creating its business strategy, while Defendant OpenAI Startup Fund GP I, LLC is the general partner of Defendant OpenAI Startup Fund I, LP, responsible for managing and operating the day-to-day business affairs of Defendant OpenAI Startup Fund I, LP.

80. Defendant OpenAI, LP directly infringed upon Plaintiffs' and Class members' copyrighted works through the unauthorized use and reproduction of the works, and preparation of derivative works by ChatGPT. As discussed above, Plaintiffs' and Class members' protected works were used to train GPT models. Because the GPT models are based on expressive information extracted from Plaintiffs' and Class members' works, Defendant OpenAI, LP is directly liable for unauthorized use, reproduction, display of copyrighted works, as well as creation of derivative works through ChatGPT's responses. Therefore, Defendant OpenAI, LP directly infringed upon Plaintiffs' and Class members' exclusive rights under 17 U.S.C. § 106.

81. Defendants OpenAI, Inc., OpenAI OpCo, LLC, OpenAI GP, LLC, OpenAI Startup Fund GP I, LLC, OpenAI Startup Fund I, LP, and OpenAI Startup Management LLC are vicariously liable for the infringement alleged herein because they had the right and ability to supervise and control the infringing activity but failed to stop the infringing conduct.

82. Furthermore, Defendants have a direct financial interest in the infringing conduct and received revenue in connection with the development, deployment, and advancement of the GPT models and ChatGPT. Each entity profited from the advancement of GPT models and ChatGPT.

83. These committed acts of copyright infringement were willful, intentional, and malicious and thus subjects Defendants to liability for statutory damages under Section 504(c)(2) of the Copyright Act of up to $150,000 per infringement.

84. Plaintiffs and Class members have been injured by Defendants' acts of vicarious copyright infringement. Plaintiffs and the Class are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law.

THIRD CAUSE OF ACTION

DIGITAL MILLENNIUM COPYRIGHT ACT – REMOVAL OF COPYRIGHT MANAGEMENT INFORMATION 17 U.S.C. § 1202(B)

85. Plaintiffs incorporate by reference all allegations of the preceding paragraphs as though fully set forth herein.

86. Plaintiffs bring this claim on behalf of herself and on behalf of the Class against Defendants.

87. Plaintiffs and Class members included one or more forms of copyright-management information in each of Plaintiffs' and Class members' infringed works, including: copyright notice, title and other identifying information, the name or other identifying information about the owners of each book, terms and conditions of use, and identifying numbers or symbols referring to the copyright-management information.

88. Without the authority of Plaintiffs and the Class, OpenAI copied Plaintiffs' and Class members' works and used them as training data for its GPT software. By design, the training process does not preserve any copyright-management information. Therefore, OpenAI intentionally removed copyright-management information from Plaintiffs' and Class members' works in violation of 17 U.S.C. § 1202(b)(1).

89. OpenAI's removal or alteration of Plaintiffs' and the Class's copyright-management information has been done knowingly and with the intent to induce, enable, facilitate, or conceal infringement of Plaintiffs' and the Class's copyrights.

90. Without the authority of Plaintiffs and the Class, Defendants created derivative works based on Plaintiffs' and Class members' works. By distributing these works without their copyright-management information, OpenAI violated 17 U.S.C. § 1202(b)(3).

91. OpenAI knew or had reasonable grounds to know that this removal of copyright-management information would facilitate copyright infringement by concealing the fact that every output from ChatGPT is an infringing derivative work, synthesized entirely from expressive information found in the training data.

92. Plaintiffs and the Class have been injured by OpenAI's removal of copyright-management information. Plaintiffs and the Class are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law, including full costs and attorneys' fees.

FOURTH CAUSE OF ACTION

VIOLATIONS OF THE CALIFORNIA UNFAIR COMPETITION LAW CAL. BUS. & PROF. CODE §§ 17200, *ET SEQ.*

93. Plaintiffs and the Class incorporate by reference each preceding and succeeding paragraph as though fully set forth at length herein.

94. Plaintiffs bring this claim on behalf of herself and on behalf of the Class against Defendants.

95. The California Unfair Competition Law ("UCL") prohibits acts of "unfair competition," including any "unlawful, unfair or fraudulent business act or practice" and "unfair, deceptive, untrue or misleading advertising." Cal. Bus. & Prof. Code § 17200.

96. Defendants have engaged in unfair competition and unfair, unlawful or fraudulent business practices by the conduct, statements, and omissions described above because it illegally collected and used Plaintiffs' and the Class's copyrighted works to train its GPT models.

97. The unlawful business practices described herein violate the UCL because Defendants used Plaintiffs' and the Class's protected works to train its GPT software for Defendants' own commercial profit without Plaintiffs' and the Class's authorization. Defendants further knowingly designed ChatGPT to include portions or summaries of Plaintiffs' copyrighted works without attributions in its responses, and they unfairly profit from and take credit for developing a commercial product based on unattributed reproductions of those stolen writing and ideas.

98. These acts and practices have deceived Plaintiffs and are likely to deceive the public into believing that Plaintiffs and the Class have granted OpenAI the right to use its copyrighted materials. In failing to disclose the sources of its training datasets and suppressing other material facts from Plaintiffs and Class members as well as the public, Defendant breached its duties to disclose these facts, violated the UCL, and caused injuries to Plaintiffs and Class members. The omissions and acts of concealment by Defendants pertained to information that was material to Plaintiffs and Class members, as it would have been to all reasonable consumers.

99. The injuries suffered by Plaintiffs and Class members are not greatly outweighed by any potential countervailing benefit to consumers or to competition, nor are they injuries that Plaintiffs and Class members should have reasonably avoided.

100. Defendant's acts and practices are unlawful because they violate California Civil Code §§ 1668, 1709, 1710, and 1750 *et seq.*, and California Commercial Code § 2313.

101. Plaintiffs seek to enjoin further unlawful, unfair and/or fraudulent acts or practices by Defendants, to obtain restitutionary disgorgement of all monies and revenues

generated as a result of such practices, and all other relief allowed under California Business & Professions Code § 17200.

FIFTH CAUSE OF ACTION

NEGLIGENCE

102. Plaintiffs incorporate by reference the allegations of all foregoing paragraphs as if they had been set forth in full herein.

103. Plaintiffs bring this claim on behalf of themselves and on behalf of the Class against Defendants.

104. Defendants owed a duty of care toward Plaintiffs and the Class in (1) obtaining data to train its GPT models and (2) not using Plaintiffs' and the Class's protected works to train its GPT models.

105. Defendants have a common law duty to prevent foreseeable harm to others, including Plaintiffs and members of the Class, who were foreseeable and probable victims of Defendants' unlawful practices.

106. Defendants breached their duty to exercise due care by negligently, carelessly, and recklessly collecting, maintaining, and controlling Plaintiffs' and Class members' works and engineering, designing, maintaining, and controlling systems—including ChatGPT—that are trained on Plaintiffs' and Class members' works without their authorization.

107. The damages suffered by Plaintiffs and the Class were the direct and reasonably foreseeable result of Defendants' negligent breach of their duties to adequately design, implement, and maintain reasonable practices to avoid infringing protected works without consent of copyright holders.

108. Defendants' negligence directly caused significant harm to Plaintiffs and the Class.

SIXTH CAUSE OF ACTION

UNJUST ENRICHMENT

109. Plaintiffs incorporate by reference all allegations of the preceding paragraphs as though fully set forth herein.

110. By virtue of the unlawful, unfair, and deceptive conduct alleged herein, Defendants knowingly realized substantial revenue from the use of Plaintiffs' and Class members' copyrighted works for the commercial training of its GPT models used to power its ChatGPT product.

111. Defendant knew or should have known that Plaintiffs and the Class have invested substantial time and energy creating the works in which they hold a copyright.

112. Defendants were conferred significant benefits when they downloaded and copied Plaintiffs' and the Class's copyrighted works to train their GPT software without Plaintiffs' and the Class's permission. Defendant knowingly and willingly accepted and enjoyed those benefits.

113. By using Plaintiffs' and the Class's copyrighted works to train ChatGPT, Defendants caused Plaintiffs and the Class to suffer actual damages from the deprivation of the benefits of their work, including monetary damages.

114. Defendants derived profit and other economic benefits from the use of Plaintiffs' and the Class's copyrighted works to train ChatGPT.

115. It would be inequitable and unjust to permit Defendants to retain the enormous economic benefits it has obtained from and/or at the expense of Plaintiffs and Class members.

116. As a direct and proximate cause of Defendants' unjust enrichment, Plaintiffs and the Class are entitled to restitution, attorneys' fees, costs and interest.

117. Defendants' conduct is causing and, unless enjoined and restrained by this Court, will continue to cause Plaintiffs and the Class irreparable injury that cannot be compensated or measured in money.

REQUEST FOR RELIEF

WHEREFORE, Plaintiffs, individually and on behalf of members of the Class defined above, respectfully request that the Court enter judgment against Defendants and award the following relief:

A. Certification of this action as a class action pursuant to Rule 23 of the Federal Rules of Civil Procedure, declaring Plaintiffs as the representative of the Class, and Plaintiffs'

counsel as counsel for the Class;

B. An order awarding declaratory relief and temporarily and permanently enjoining Defendant from continuing the unlawful and unfair business practices alleged in this Complaint and to ensure that all applicable information set forth in 17 U.S.C. § 1203(b)(1) is included when appropriate;

C. An award of statutory and other damages under 17 U.S.C. § 504 for violations of the copyrights of Plaintiff and the Class by Defendants.

D. An award of statutory damages under 17 U.S.C. § 1203(b)(3) and 17 U.S.C. § 1203(c)(3), or in the alternative, an award of actual damages and any additional profits under 17 U.S.C. § 1203(c)(2);

E. A declaration that Defendant is financially responsible for all Class notice and the administration of Class relief;

F. An order awarding any applicable statutory and civil penalties;

G. An order requiring Defendant to pay both pre- and post-judgment interest on any amounts awarded;

H. An award of costs, expenses, and attorneys' fees as permitted by law; and

I. Such other or further relief as the Court may deem appropriate, just, and equitable.

/ /

/ /

/ /

/ /

/ /

/ /

/ /

/ /

/ /

DEMAND FOR JURY TRIAL

Plaintiffs hereby demand a jury trial for all claims so triable.

DATED: September 8, 2023

Respectfully submitted,

/s/ *Daniel J. Muller*
DANIEL J. MULLER, SBN 193396
dmuller@venturahersey.com
VENTURA HERSEY & MULLER, LLP
1506 Hamilton Avenue
San Jose, California 95125
Telephone: (408) 512-3022
Facsimile: (408) 512-3023
dmuller@venturahersey.com

/s/ *Bryan L. Clobes*
Bryan L. Clobes (*pro hav vice anticipated*)
CAFFERTY CLOBES MERIWETHER & SPRENGEL LLP
205 N. Monroe Street
Media, PA 19063
Tel: 215-864-2800
bclobes@caffertyclobes.com

Alexander J. Sweatman (*pro hav vice anticipated*)
CAFFERTY CLOBES MERIWETHER & SPRENGEL LLP
135 South LaSalle Street, Suite 3210
Chicago, IL 60603
Tel: 312-782-4880
asweatman@caffertyclobes.com

Attorneys for Plaintiffs